从下面几个方面系统聊下这个问题：

1.为什么需要置信区间？

2.什么是置信区间和置信水平？

3.如何计算置信区间？

> **1.**为什么需要置信区间？

历史上最早的科学家曾经不承认实验可以有误差，认为所有的测量都必须是精确的，把任何误差都归于错误。后来人们才慢慢意识到误差永远存在，而且不可避免。即使实验条件再精确也无法完全避免随机干扰的影响，所以做科学实验往往要测量多次，用取平均值之类的统计手段去得出结果。

多次测量，是一个排除偶然因素的好办法。国足输掉比赛之后经常抱怨偶然因素，有时候是因为裁判不公，有时候是因为主力不在，有时候是因为不适应客场气候，关键是如果你经常输球，我们还是可以得出你是个弱队的结论。

而国际足联的世界排名，是根据各国球队多次比赛的成绩采用加权平均的办法统计出来的，这个排名比一两次比赛的胜负，甚至世界杯赛的名次更能说明球队的实力。但即便如此，我们也不能说国际足联的排名就是各个球队的“真实实力”。这是因为各队毕竟只进行了有限次数的比赛，再好的统计手段，也不可能把所有的偶然因素全部排出。

所以，在科学实验中总是会在测量结果上加一个误差范围。比如经过测量马云的智商是100，测量误差是±5。

这句话的意思是说，马云智商是100，但其中有正负5的统计误差，所以马云的智商范围就是[100-5,100+5]这么一个范围。

真实的智商值当然只有一个，但是这个数是多少，我们不知道，它可以是这个误差范围内的任何一个数字。

考试成绩也如此，假设一个同学考了两次才过英语四级，第一次53分，第二次63分。他说这是略有进步，我说你这不叫进步，叫都在测量误差范围之内。

在股票市场经常会看到有人为了短期的股价上涨而兴奋不已，却又对短期的股价下跌彻夜难眠。其实这都是因为不理解误差范围导致的。

想想，如果这些人真的具备了误差的概率，就会忽略误差范围内的任何波动。如果你投资的这家公司在未来10年有足够的成长空间，那么你就会忽略掉这10年期间它股价暂时的波动，因为你看到的是长期，只要长期在你预期的误差范围内就可以接受。

这里的误差范围（区间）在统计概率中就叫做置信区间。简单来说，置信区间就是误差范围。

> **2.**什么是置信区间和置信水平？

在之前我在“统计概率与投资”的课程中有讲到过到如何用样本估计总体。社群会员就问了我一个问题：在抽样调查中，样本能在多大程度上代表总体？有没有公式来表示？

统计概率思维与投资

1 什么是置信区间？

猴子的主题

来自 统计概率思维与投资

提问

张伟松　　猴子

在抽样调查中，样本能在多大程度上代表总体？有没有公式能够表示？

猴子：

可以看下中心极限定理，和置信区间这两块内容

其实这个问题的本质就是想知道数据统计的误差范围是多少。在统计概率中有个专门的名称来表示误差范围，叫置信区间。

比如我用一定量的样本数据估计出全体知乎用户的平均年龄为28岁。

如果你收集了另外一组样本，其平均年龄为35岁，是否能判断我前面的估计是错误的呢？

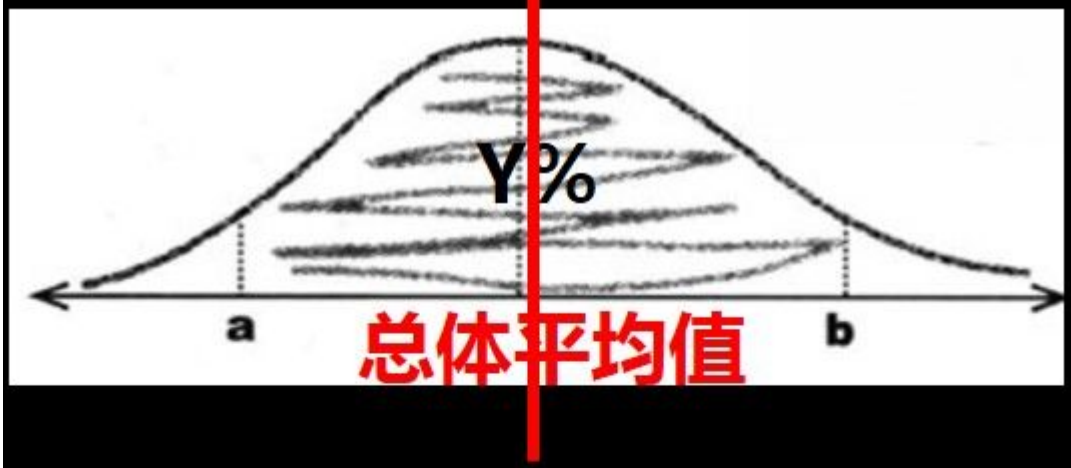
因为我们没办法知道总体平均数的真实数值，所以，我们需要给出一个误差范围来描述这个估计的准确程度。

如果你已经知道什么是中心极限定理（猴子：怎样理解和区分中心极限定理与大数定律？），就会知道：样本围绕在总体平均值周围呈现正态分布。所以下图中中间红色线是总体平均值。

（如果不懂正态分布，看这里：猴子：怎样用通俗易懂的文字解释正态分布及其意义？）

置信区间(误差范围): [a,b]

置信水平 Y%：
P(a<均值<b)=Y%



我们用中括号[a,b]表示样本估计总体平均值的误差范围的区间，由于a和b的确切数值取决于你希望自己对于“该区间包含总体均值”这一结果具有的可信程度，因此，[a,b]被称为置信区间。

同时，我们选择这个置信区间，目的是为了为了让“a和b之间包含总体平均值”这一结果具有特定的概率，这个概率就是置信水平。

假设我设定的置信水平是95%，也就是说如果我做100次抽样，会有95个置信区间包含了总体平均值。

> **3.**如何计算置信区间？

其实，任何的统计概率知识都没有那么高大上，同样的，计算置信区间也是一种套路。如果你学会学会下面我介绍的计算置信区间的4个步骤，你也可以轻松计算出置信水平。

第1步：确定要求解的问题是什么

假设我是医院的数据分析师，想知道新药物A对神经的反应时间。因此，需要为总体平均值构建一个置信区间。这决定了我需要抽取一个合适的样本。通过样本的数据来估计出总体的数据

第2步：求样本的平均值和标准误差

当样本大小大于30时，抽取的样本符合中心极限定理。

为了应用中心极限定理，我们后面所指的样本大小都是大于30的。

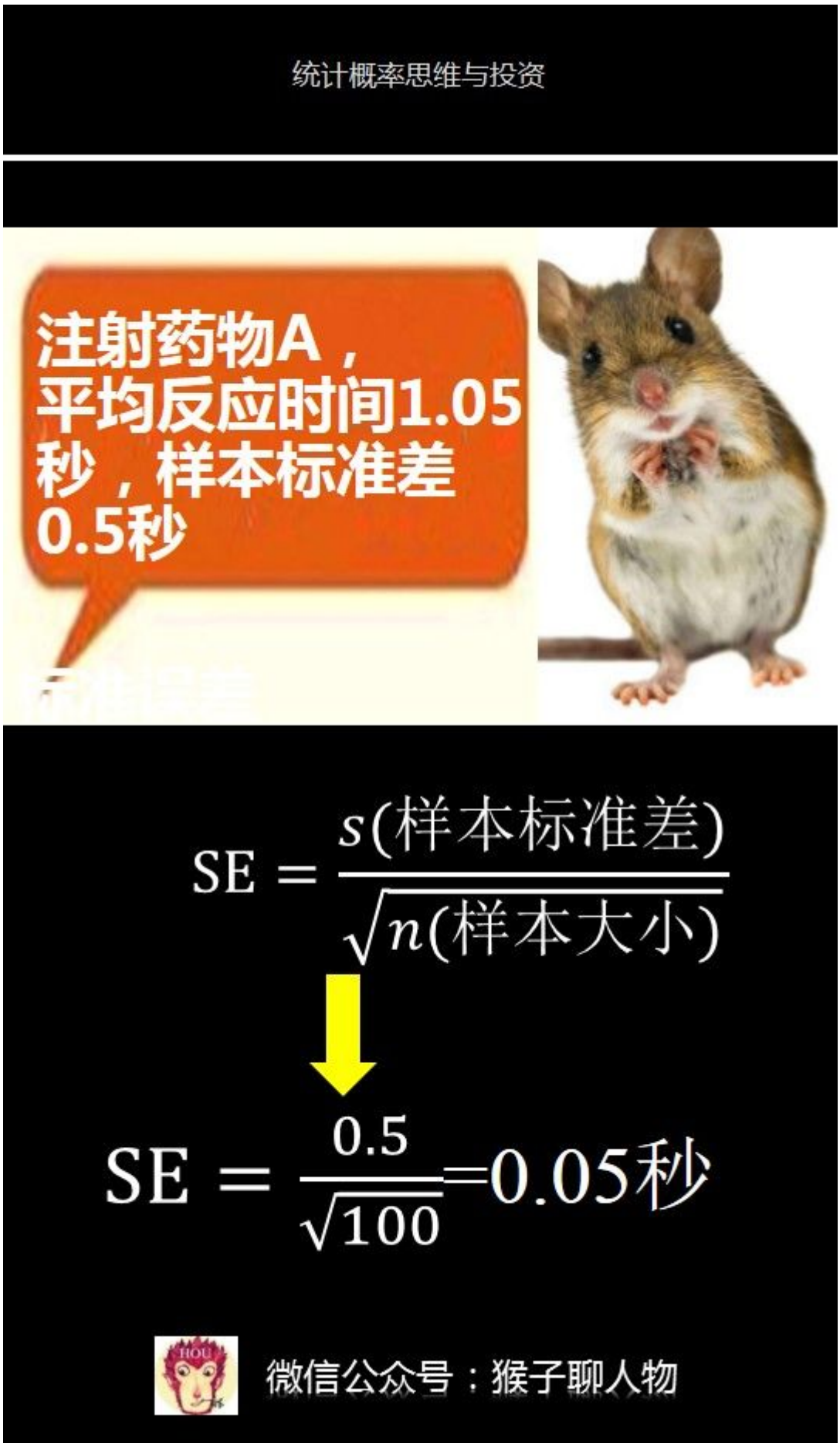
为了用样本估计出总体的平均值，也就是新药对神经的平均反应时间。我找来100只老鼠作为样本来做实验，对每只老鼠都注射了药物A，对其进行神经刺激，并记录反应时间。最后得到平均反应时间是1.05秒。样本标准差是0.5秒。

根据中心极限定理，我可以用样本平均值估计出总体平均值也是1.05秒。

当我兴高采烈的把这个结果告诉老板，老板为了验证我数据的准确性，又找人重复了我的实验，发现样本的平均反应时间是1秒。发现与我给的数据不一样，是不是我的数据出错了呢？

其实，是我一开始给老板的数据信息是不准确的，没有给出数据的误差范围。为了计算出误差范围，我需要先计算出标准误差。

标准误差SE等于样本标准差除以n的开方。最后算出标准误差等于0.05秒。


统计概率思维与投资
注射药物A，
平均反应时间1.05
秒，样本标准差
0.5秒
$SE = \frac{s(样本标准差)}{\sqrt{n(样本大小)}}$
$SE = \frac{0.5}{\sqrt{100}} = 0.05秒$
微信公众号：猴子聊人物

第3步：确定置信水平

那么由谁来决定置信水平？多大的置信水平才合适？

答案完全取决于你的具体情况以及你需要对“区间中包含总体平均值”这一说法有多大信心。

关键是记住一点：置信水平越高，区间越宽，置信区间包含总体平均值统计量的概率越大。

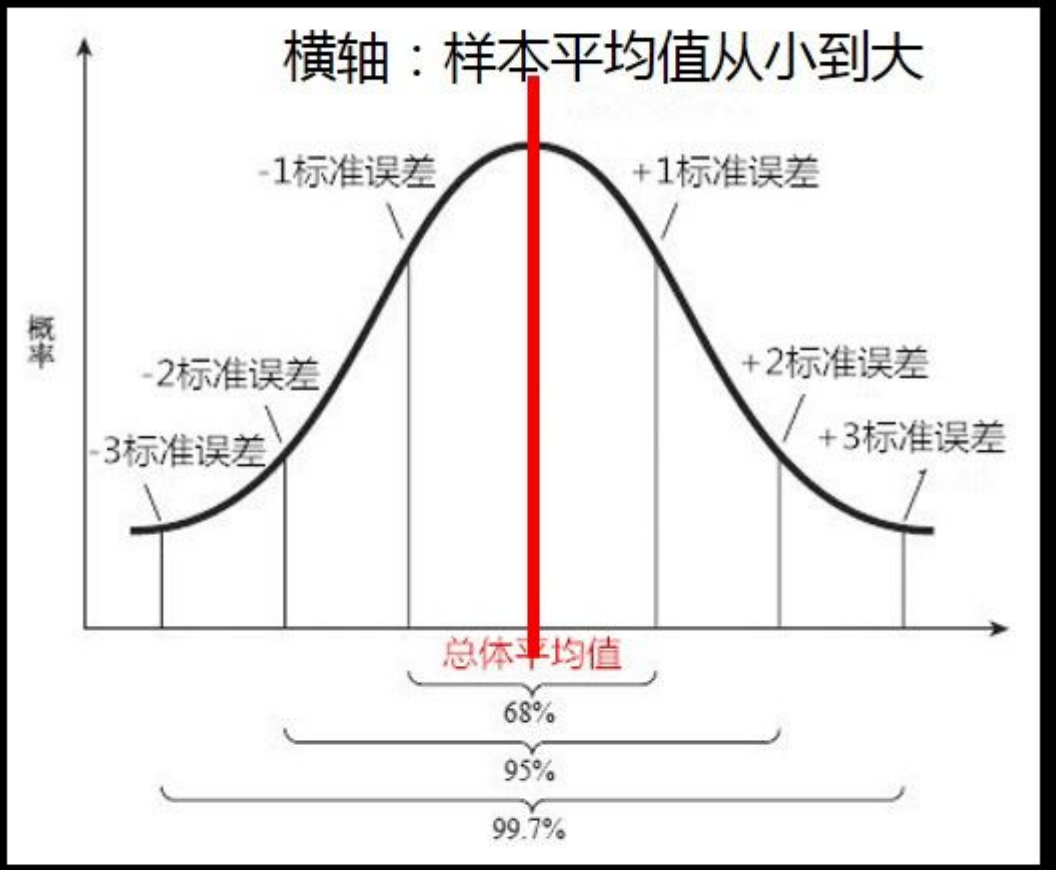
常用的置信水平是95%。其实，这个数字并不是必然的，而是人为设定的。

那么置信区间为什么通常是95%呢？

下面图中是中心极限定理的样本平均值概率图。这个图在后面一直会用到，这里再重点介绍下这个图。

第3步：确定置信水平

常用的置信水平95%



样本平均值 概率分布图

微信公众号：猴子聊人物

图中横轴是样本平均值从小到大，纵轴是样本平均值对应的概率。根据中心极限定理，我们知道不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的平均值周围，并且呈正态分布。

所以图中的中间位置红色线是总体平均值。

而有95%的样本均值会落在2个标准误差范围内，这也是为什么会选择95%作为置信区间的原因。

（置信水平的设定是有影响的——如果我们对置信水平要求过高，我们可能会拒绝实际上是正确的理论（犯了I类错误）；

如果我们对置信水平要求过低，我们可能会接受错误的理论（犯了II类错误）。

并没有一个万全之策能够让犯两种错误的可能性同时降低，我们必须做出选择。鉴于我们更加不喜欢犯II类错误，所以我们习惯于把置信水平设置在高水平。）

第4步：求出置信区间上下限的值

现在我们来求置信区间[a,b]的上限a和下限b的值。

我们如果能计算出a离总体平均值多少个标准误差，那么我们就可以知道a的值了。为什么这么说呢？

假设a离总体平均值2个标准误差，那么a=总体平均值-2个标准误差

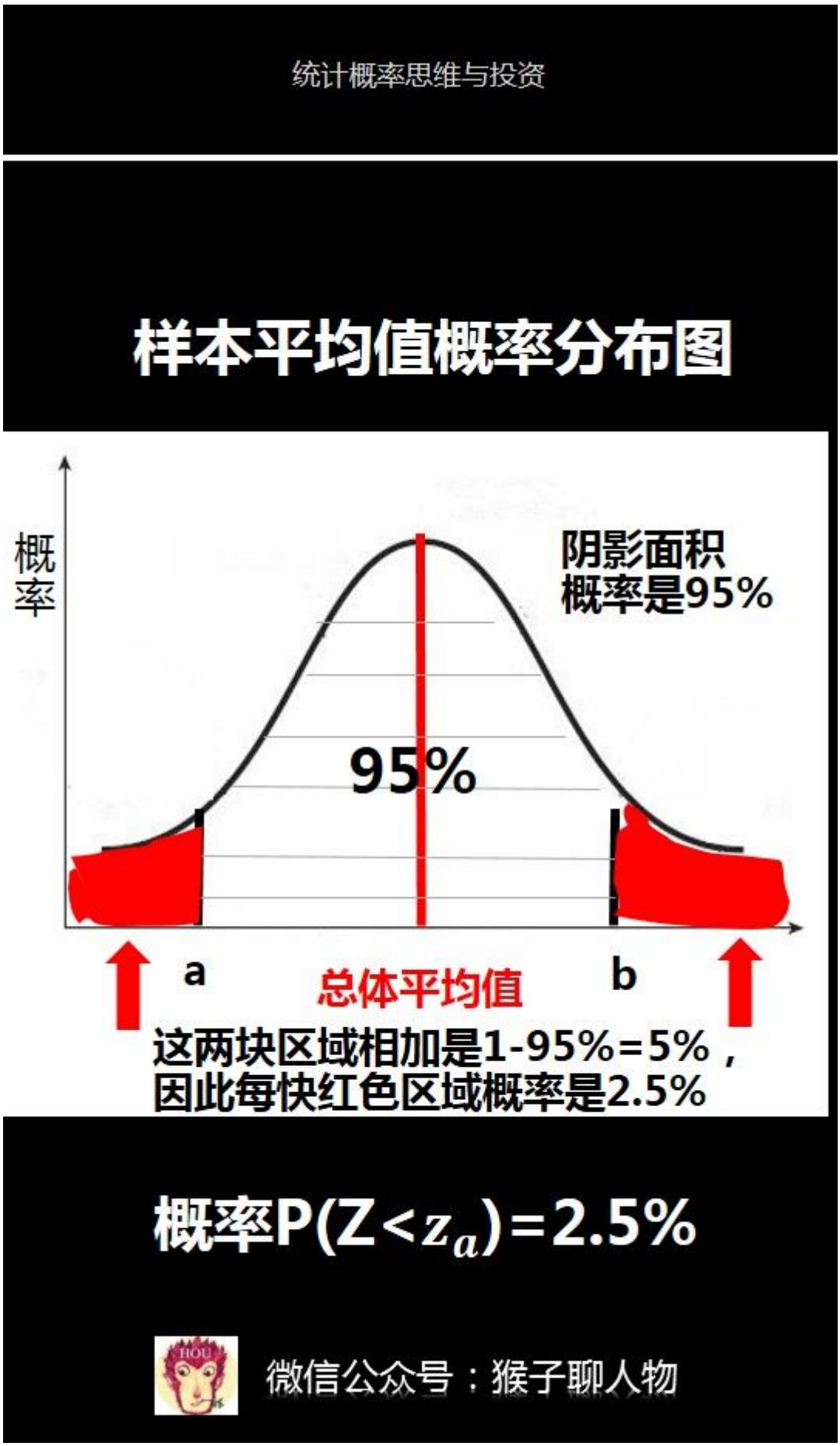
同样的，根据根据正态分布的对称性，我们就可以知道b的值，也就是b=总体平均值+2个标准误差

这里距离平均值几个标准误差，就是我之前聊过的标准分（猴子：有了方差为什么需要标准差？）。所以，现在问题变的很简单了，只要我们求出a对应的标准分是多少就可以了。

我们用Z来表示几个标准误差，就是Z乘以标准误差。下面我们看下如何计算出标准分z的值。

现在我们知道，下图中阴影部分，也就是置信区间a和b包括的概率是置信水平95%，

由于整个概率的和是1，所以我们可以知道图中两块红色区域的概率相加是1-95%=5%，而两端是对称的，所以每块红色区域的概率是2.5%



也就是概率P(Z<Za)=2.5%，现在知道概率了，我们可以根据z表格来查询获取到对应的z值。

z表格也叫标准正态分布表，它是标准正态分布中，标准分与概率数值的对应关系表。这个表格就是在你知道表标准分的情况下，可以快速查找到对应的概率值。

同样的反过来，你知道概率值，也可以查找到对应的标准分z是多少。

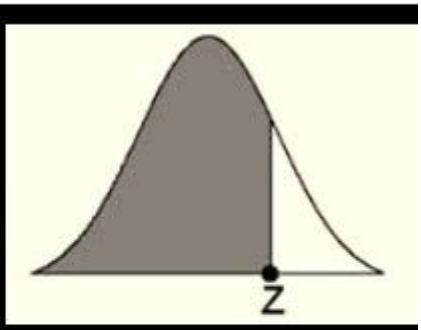
现在我们已经知道了概率值是2.5%，那么就是查找对应的标准分z是多少呢？

在表格中我们查找到概率值2.5%对应的最左边第一列的值是-1.9，对应的最上边第一行的值是0.06。

根据Z表格给出的是小于标准分z的概率，也就是p（Z<z）。查找概率时，需要用第一列和第一行找出数值Z，在表格中，z数值的第一位小数值在表格最左边的第一列。z数值的第2位小数值在表格的第一行。所以z=-1.96

统计概率思维与投资

P(Z<z)
等于这块阴影面积



概率P(Z<z)=2.5%

0.06

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0	
-1.9							
		0.0351	0.0344	0.0336	0.0329	0	

0.025

-1.9

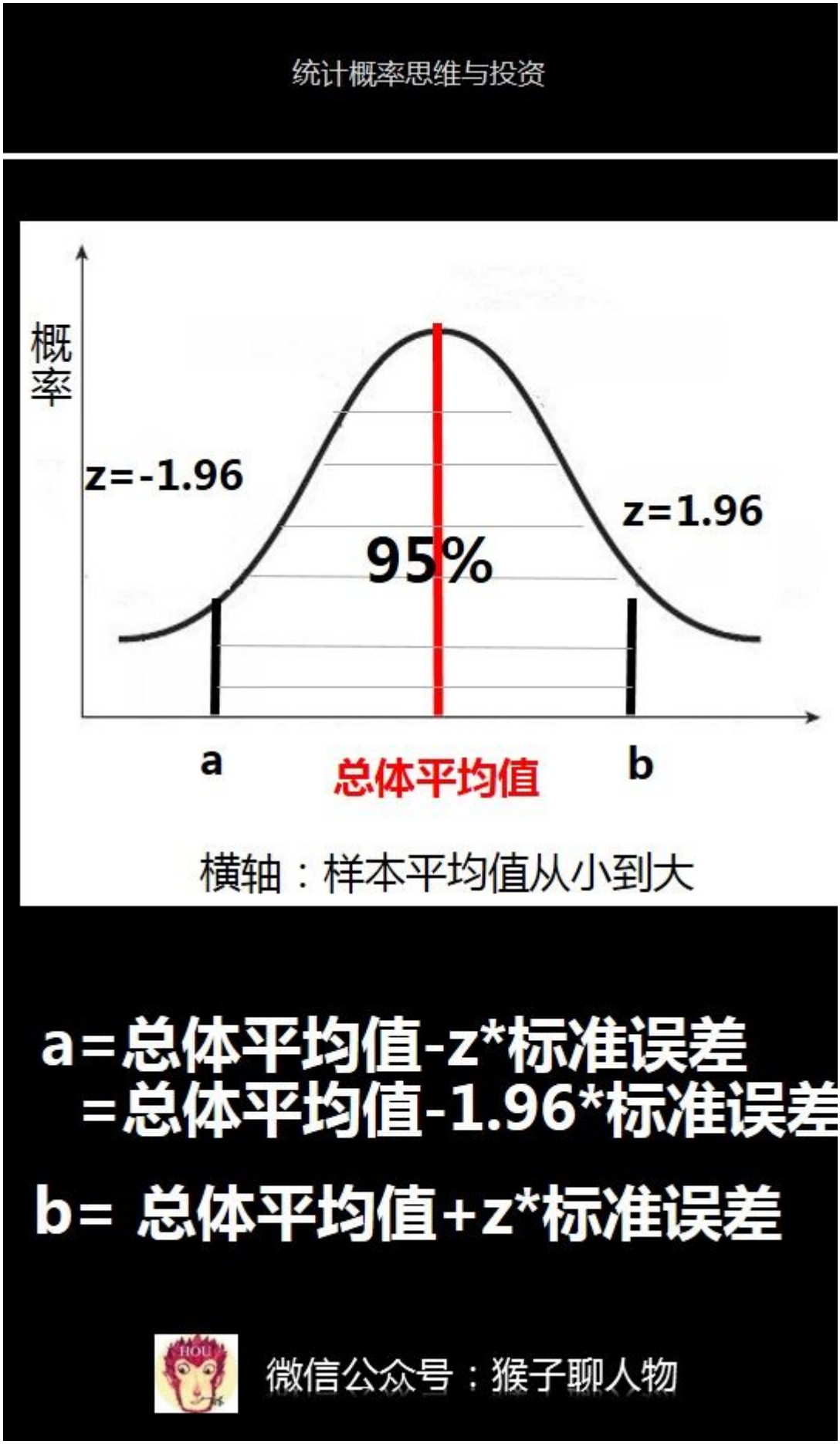
标准分z=-1.96

微信公众号：猴子聊人物

Z=-1.96表示距离总体平均值左边1.96个标准误差，所以是负数。而b在总体平均值右边，所以z是正数，也是1.96个标准误差。所以，这里的z就是1.96：

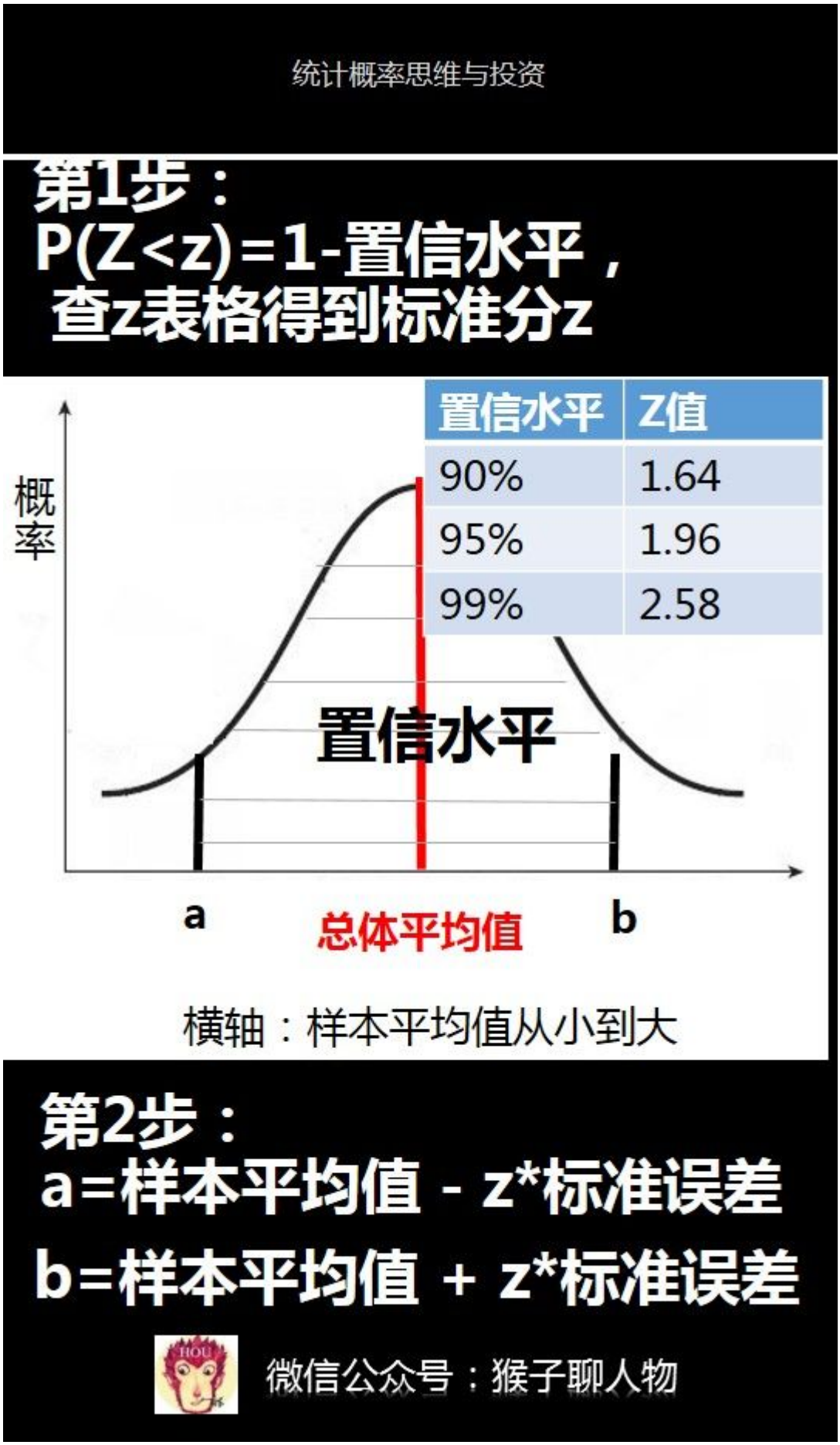
a=总体平均值**-1.96***标准误差

b=总体平均值**+1.96***标准误差



而之前我们已经求得标准误差，那么总体平均值是多少呢？

根据中心极限定理，样本平均值约等于总体平均值，所以我们可以得到下面图片中置信区间的一般表达方式。



我们总结下前面计算的过程，你就更容易理解了。

第1步，我们根据置信水平，知道了概率值，并查找z表格得到了对应的z值

其实常用的置信水平对应的z值我已经放在图中了，你直接就可以套用。比如置信水平90%对应的z值是1.64,95%的置信水平对应的z值是1.96

第2步，我们计算a和b 的值

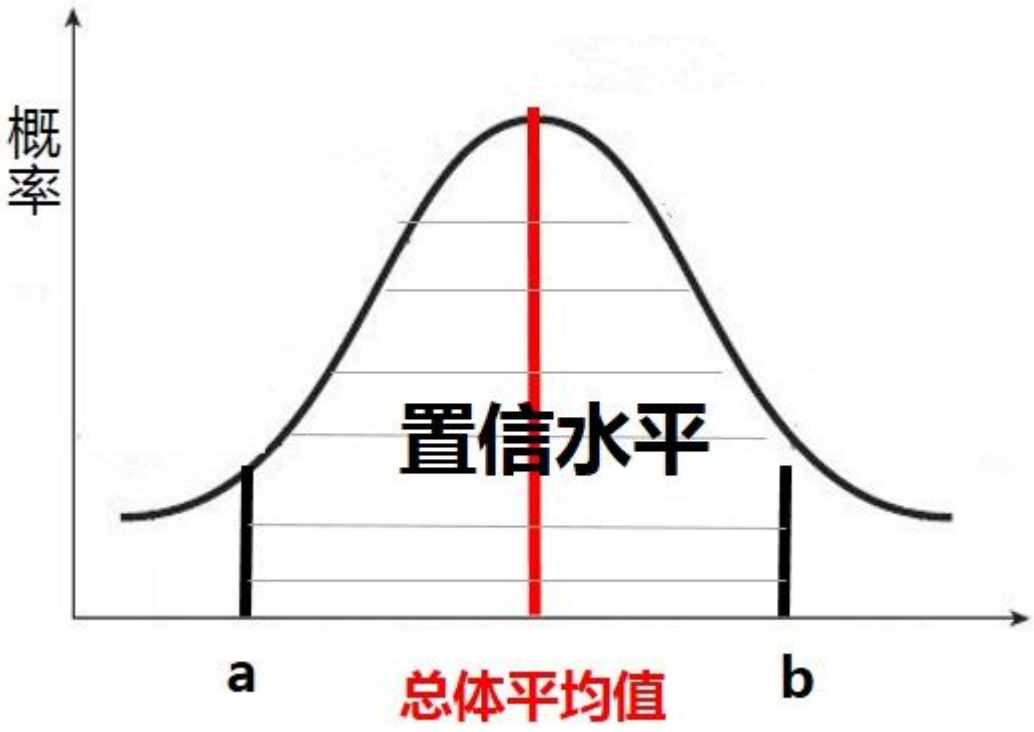
a=样本平均值-z*标准误差

b=样本平均值+z*标准误差

既然只要在简便算法中带入数值就行，为什么讲那么多步骤呢？

讲这些步骤是为了让你看清楚问题实质，理解置信区间的构建过程。大多数时候，你只要带入数值就行了。

下面图片我们将这个置信区间的公式带入我们前面老鼠实验药物的例子中，就可以得到下图中的置信区间：



横轴：样本平均值从小到大

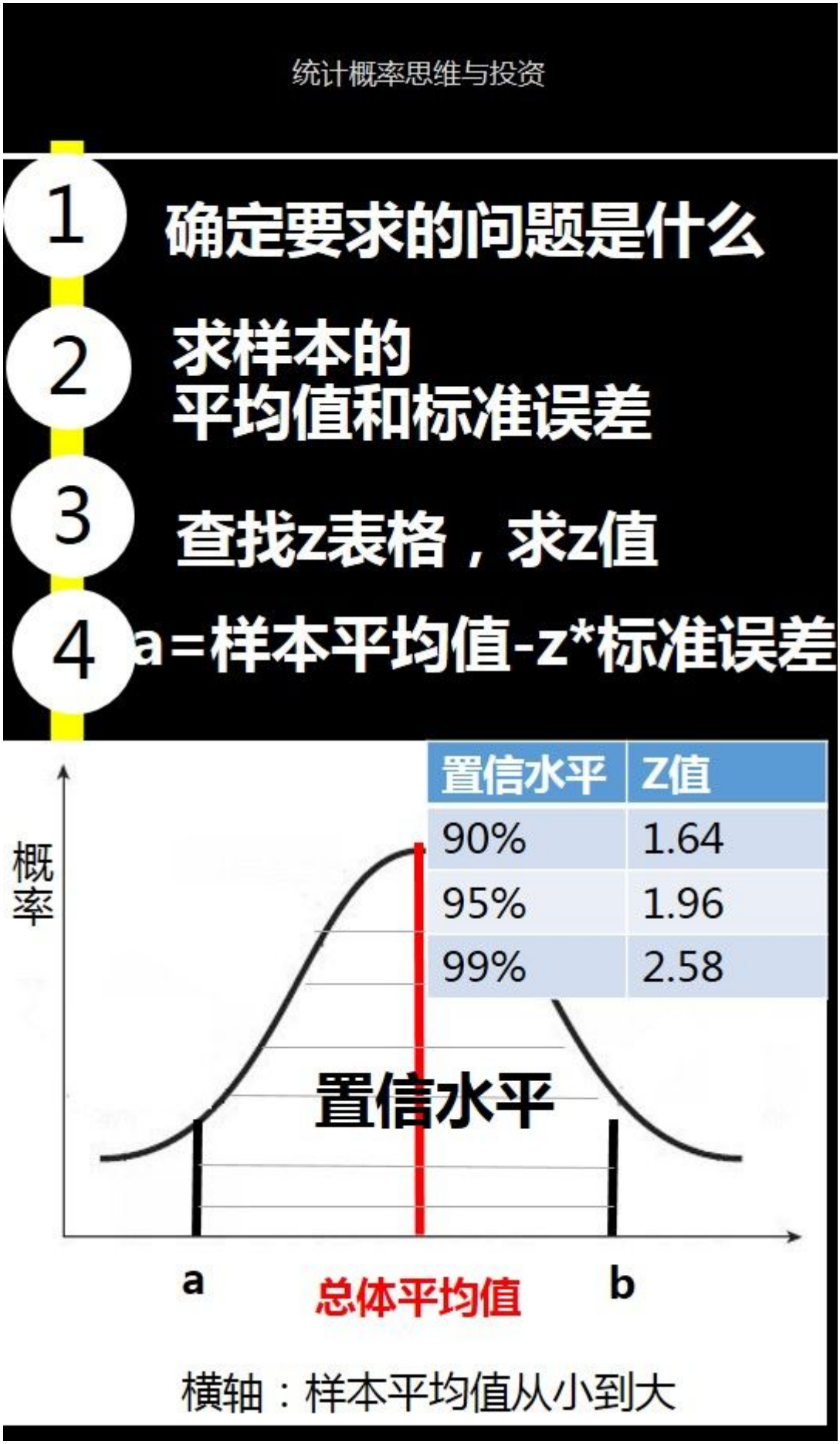
a=样本平均值 - z*标准误差
=1.05-1.96*0.05
= -0.952

b=样本平均值 + z*标准误差
= 1.148

置信水平95%，
置信区间[-0.952,1.148]

微信公众号：猴子聊人物

前面我已经详细解释了计算置信区间的4个步骤，你也已经理解了。现在我们来总结下计算置信区间的4个步骤，你会发现这比你想象中简单很多。



1. 确定要求解的问题是什么

比如我们想要通过样本来估计总体的平均值

2.求样本的平均值和标准误差

第3步：确定置信水平

常用的置信水平是95%，因为这样可以保证样本的平均值会落在总体平均值2个标准误差范围内

3.查找z表格，求z值

如果你的置信水平是图中的95%，可以直接获取到对应的z值

4.计算置信区间

a=样本平均值- z*标准误差

b=样本平均值+ z*标准误差

4. 一句话总结前面的知识

如果你看统计概率方面的书，很多书中也会有讲T分布下的置信区间计算，也就是当样本数量小于30时，样本分布符合T分布。这里我不准备聊这个知识，因为太多会让你大脑内存溢出。

你只需要记住有这么个T分布，当你拿到的数据样本不足30时，才会用到它。

大部分情况下，我们是可以获取到大于30的样本，这时候样本平均值是符合正态分布的，用我聊的步骤来计算就可以了。

统计概率思维与投资

序号	知识点	和猴子用一句话记住
1	**置信区间**	误差范围
2	**置信水平**	置信水平目的是为了为了让“a和b之间包含总体平均值”这一结果具有特定的概率，这个概率就是置信水平。假设我设定的置信水平是95%，也就是说如果我做100次抽样，会有95个置信区间包含了总体平均值。
3	**如何计算置信水平**	a=样本平均值 - z*标准误差 b=样本平均值 + z*标准误差

置信水平	Z值
90%	1.64
95%	1.96
99%	2.58

微信公众号：猴子聊人物